DOCUMENT RESUME

ED 478 203 TM 035 049

AUTHOR Briggs, Derek. C.

TITLE Causal Inference and the Heckman Model.

PUB DATE 2003-04-00

NOTE 57p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, April 22-24, 2003).

PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)

EDRS PRICE EDRS Price MF01/PC03 Plus Postage.

DESCRIPTORS *Causal Models; College Entrance Examinations; *Program Effectiveness; Regression (Statistics); Selection; Statistical Bias; *Statistical Inference; *Test Coaching

IDENTIFIERS Heckman (J J)

ABSTRACT

In the social sciences, evaluating the effectiveness of a program or intervention often leads researchers to draw causal inferences from observational research designs. Bias in estimated causal effects becomes an obvious problem in such settings. This paper presents the Heckman Model as an approach sometimes applied to observational data for the purpose of estimating an unbiased causal effect. The paper shows how the Heckman model can be viewed as an extension of the linear regression model, and discusses in some detail the assumptions necessary before either approach can be used to make causal inferences. Linear regression and the Heckman Model can make different assumptions about the relationship between two equations in an underlying behavioral model: a response schedule and a selection function. Under linear regression the two equations are assumed to be independent; under the Heckman Model, the two equations are allowed to be correlated. The Heckman Model is particularly sensitive to the choice of variables included in the selection function. This is demonstrated empirically in the context of estimating the effect of commercial coaching programs on the Scholastic Assessment Test (SAT) performance of high school students. Coaching effects are estimated for both sections of the SAT using data from the National Education Longitudinal Study of 1988. Small changes in the selection function are shown to have a big impact on estimated coaching effects under the Heckman Model. (Contains 2 tables, 8 figures, and 42 references.) (Author/SLD)

Running head:

CAUSAL INFERENCE AND THE HECKMAN MODEL

Causal Inference and the Heckman Model

Derek C. Briggs

University of California, Berkeley

April, 2003

Paper presented at the

National Council on Measurement in Education Annual Meeting, Chicago, IL,

Abstract

In the social sciences, evaluating the effectiveness of a program or intervention often leads researchers to draw causal inferences from observational research designs. Bias in estimated causal effects becomes an obvious problem in such settings. I present the Heckman Model as an approach sometimes applied to observational data for the purpose of estimating an unbiased causal effect. I show how the Heckman Model can be viewed as an extension of the linear regression model, and discuss in some detail the assumptions necessary before either approach can be used to make causal inferences. Linear regression and the Heckman Model make different assumptions about the relationship between two equations in an underlying behavioral model: a response schedule and a selection function. Under linear regression the two equations are assumed to be independent; under the Heckman Model, the two equations are allowed to be correlated. The Heckman Model is particularly sensitive to the choice of variables included in the selection function. This is demonstrated empirically in the context of estimating the effect of commercial coaching programs on the SAT performance of high school students. I estimate coaching effects for both sections of the SAT using data from the National Education Longitudinal Study of 1988 (NELS). Small changes in the selection function are shown to have a big impact on estimated coaching effects under the Heckman Model.

Introduction

In the social sciences, evaluating the effectiveness of a program or intervention often leads researchers to draw causal inferences from observational research designs. Suppose a study involves a sample of high school students. One group of students takes part in a program that promises to improve their type-writing speed, and the other group does not. After the former group has completed the program, the average number of words typed per minute for the two groups are compared. Can a difference between the groups be attributed to the program? This is the key question in making causal inferences. In this hypothetical example, treatment and control groups are not randomly assigned. Thus, outcome differences between the groups may be explained by other characteristics on which the two groups differ. Causal effect estimates calculated by comparing averages will tend to suffer from bias[1], which can lead to incorrect inferences about program effectiveness.

A number of statistical methods have been used in observational settings to control for bias. There is a common thread running through all these approaches: the idea that an observational study can be considered as a randomized experiment, conditional on certain covariates. The approaches differ in the statistical assumptions they make and the methods they apply to the data. In this paper the focus is on a method of controlling for

[1] The term bias is defined here in a statistical context (e.g. an estimated causal effect is biased), not an educational measurement context (e.g. the test items are biased against certain types of students).

bias known as the Heckman Model.[2] I present the Heckman Model as an extension of the linear regression model, and compare the similarities and differences between the two models as approaches for drawing causal inferences. While the Heckman Model is a well-established approach among econometricians, its use is less common among educational statisticians. The first part of this paper will serve as a didactic introduction to the Heckman Model for the benefit of this latter audience. The rest of the paper raises questions about the sensitivity of the Heckman Model to its specification, and is aimed at the wider audience of social scientists who might employ the approach as a tool for causal inference.

To give this presentation an applied context, both the linear regression model and Heckman Model are used to evaluate the effectiveness of coaching programs in improving performance on the SAT. The SAT is a standardized test required for admission at almost all competitive four-year colleges in the United States.[3] The test has a math and verbal section, each scored on a scale that ranges from 200 to 800 with standard deviation of about 110 points. Each year about two million high school students take the SAT at a cost of about $25 each. Coaching for the SAT (and many other

[2] Three other popular approaches that are sometimes used in this context include the Propensity Matching Model (Rosenbaum & Rubin, 1983), two stage least squares (Greene 1993, 603-10), and structural equation modeling (Jöreskog & Sörbom, 1996).

[3] As of 1994, the SAT became the SAT I. For the sake of consistency, the term SAT is used throughout generically to represent a multiple-choice test used for purposes of college admission. For a historical description of the SAT in the context of its use in college admissions decisions see Zwick, 2002; Lawrence et. al., 2001; Lemann, 1999.

standardized tests) is a multimillion dollar industry. Companies such as Kaplan and The Princeton Review charge roughly $800 for 30-40 hours of instruction, and attribute to their programs average gains of 100-140 points on the combined math and verbal sections of the test (Schwartz, 1999). Private tutors, books, videos and computer software are also available, at a price, to help students prepare for the test. It has become widely accepted among the general public that coaching has a large effect on student scores. Yet most of the published research on the topic suggests that the combined coaching effect is fairly small, in the range of about 20 to 30 points (cf. Messick, 1980; Messick & Jungeblut, 1981; Becker, 1990; Powers, 1997, Powers & Rock, 1999). One problem for this line of research has been that coaching effect estimates are usually based on studies with observational designs, making clear causal inference about coaching effectiveness elusive.

When certain assumptions hold the Heckman Model is a statistical approach that could be used to estimate an unbiased effect of coaching. On the face of things, the Heckman Model is an attractive solution to the problem of bias in an estimated coaching effect. It extends the linear regression model by turning the problem of confounding due to a latent covariate (i.e. "selection bias") into that of confounding due to a measured covariate omitted from a regression equation (i.e. "omitted variable bias"). The theoretical benefits of the approach are considerable, but as I demonstrate, there are rather large empirical costs.

Bias and Statistical Solutions

The objective in a randomized study is to determine the strength of a hypothesized causal relationship between, for example, coaching status (*COACH*) and SAT scores (*Y*) as in Figure 1.

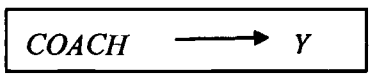


Figure 1. Causation

In an observational study with the same objective, it is usually conceivable, and often highly likely that other covariates may confound the relationship between the treatment and the outcome, as in figure 2.

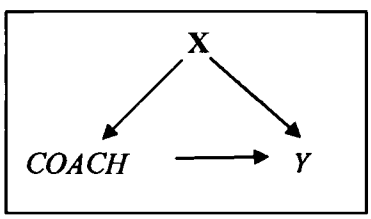


Figure 2. Confounding

In Figure 2, **X** represents a set of covariates that might include each student's pre-coaching SAT score and socioeconomic status. These covariates may influence post-coaching performance on the SAT and also be correlated with coaching status. The relationship between *Y* and *COACH* is thus confounded by **X**. A statistical approach frequently applied to correct for the possibility of bias due to confounding is linear

regression. In what follows a specialized version of linear regression is presented to facilitate a comparison with the Heckman Model.[4]

Consider the following behavioral model:

$$f_i(COACH) = a + bCOACH + \mathbf{X}_i\mathbf{c} + \sigma\varepsilon_i \quad (1)$$

$$COACH_i = 1 \Leftrightarrow \alpha + \mathbf{X}_i\gamma + \delta_i > 0. \quad (2)$$

The model consists of a *response schedule* (1) and a *selection function* (2). In the response schedule, a student's potentially observable SAT score is a function of *COACH*. Two different scores are possible for student *i*, depending on whether *COACH* = 1 or 0. The variable *COACH* is in theory manipulable—if its value is changed, the SAT score subsequently observed for student *i* will change as well (unless, of course, there is no coaching effect). The observed covariates in the vector $\mathbf{X}_i$ are fixed characteristics of each student—they cannot be manipulated by the researcher. The response schedule specified here assumes a linear relationship between the variable *COACH* and the SAT score, with a constant effect across individuals, represented by the parameter *b*. Likewise, the effect of $\mathbf{X}_i$ is linear, and $\mathbf{c}$ is the same for all students. The "error" term $\sigma\varepsilon_i$ represents the deviation of student *i*'s SAT score from its expected value. In an experimental setting, the observed value of *COACH* for student *i* would be assigned by the researcher with a known probability. Here, the observed value of *COACH* is assumed to be governed by the selection function. I describe the selection function in more detail in the context of the

[4] The specialization comes primarily from restrictions on the distribution of the unobservable error terms. Linear regression could be used to make causal inferences under more general assumptions. See for example, Freedman, 2002 and Holland, 2001.

Heckman Model. For now it suffices to note that the function implies that student i's decision to seek coaching depends on observable covariates in the vector $\mathbf{X}_i$, and on the latent covariate δ_i.

In an observational study, the researcher observes the triple $\{Y_i, COACH_i, \mathbf{X}_i\}$, where $COACH_i$ is determined by the selection function (2), and

$$Y_i = f_i(COACH_i) = a + bCOACH_i + \mathbf{X}_i\mathbf{c} + \sigma\varepsilon_i \quad (3)$$

is determined by the response schedule (1). Further statistical assumptions must be made:

i) $(\varepsilon_i, \delta_i)$ are independently and identically distributed (iid) in i with a standard normal distribution;

ii) $\{\mathbf{X}_i: i = 1, \ldots, N\}$ is independent of $\{\varepsilon_i, \delta_i: i = 1, \ldots, N\}$.

iii) δ_i and ε_i are independent within student i.

According to these assumptions, the data generated from (1) and (2) have the feature that $\{(\mathbf{X}_i, \delta_i): i = 1, \ldots, N\}$ is independent of $\{\varepsilon_i: i = 1, \ldots, N\}$. It follows therefore that $\{(\mathbf{X}_i, COACH_i): i = 1, \ldots, N\}$ is independent of $\{\varepsilon_i: i = 1, \ldots, N\}$. Thus, $COACH_i$ and $\mathbf{X}_i$ are exogenous, so ordinary least squares (OLS) can be used to get unbiased estimates for the parameters a, b and $\mathbf{c}$, by running a linear regression of Y_i on a constant, $COACH_i$ and $\mathbf{X}_i$.

In making causal inferences about the effectiveness of coaching, b is the parameter of interest, with a causal interpretation because of Equation 1. In other presentations of unbiased parameter estimation using linear regression, it is assumed that

$$E(\varepsilon_i \mid COACH_i, \mathbf{X}_i) = 0. \quad (4)$$

This follows from assumptions i, ii and iii.

The linear regression adjustment for $\mathbf{X}_i$ is essentially a replacement for random assignment in an experimental design. However, the assumptions that treatment status and covariate values are independent of the error terms, and that error terms are independent within and across cases, are clearly rather difficult to defend in the absence of a theoretical understanding of the causal mechanism at work. A common criticism among statisticians is that the plausibility of such assumptions in observational settings is seldom given adequate consideration.[5]

Implicit in estimating the effect of coaching by linear regression is that any differences between coached and uncoached students related to SAT performance are accounted for by $\mathbf{X}_i$: bias is a function of variables omitted from the regression equation. To see this more clearly, consider the linear regression equation presented in matrix format. Let $\mathbf{M}$ be a matrix containing the constant term and observed values of $COACH_i$ for $i = 1, \ldots, N$ students in a given study. Let the matrix $\mathbf{X}$ represent the collection of covariate values $\mathbf{X}_i$ for $i = 1, \ldots, N$. Similarly, the SAT score Y_i and the error term ε_i are collected into the vectors $\mathbf{Y}$ and $\boldsymbol{\varepsilon}$. Then, in matrix format

$$\mathbf{Y} = \mathbf{Mb} + \mathbf{Xc} + \boldsymbol{\varepsilon}, \tag{5}$$

[5] Some exchanges along these lines can be found in Freedman (1987; 1995). For a different interpretation of the ε term in line with the Neyman-Rubin model for causal inference, see Holland, 2001.

where $\mathbf{b} = [a\ b]$. If instead of the regression implied by Equation 5, the researcher regressed **Y** on **M**, omitting the confounding variables **X**, then the OLS estimate of the average coaching effect would be biased, since

$$\begin{aligned} \hat{\mathbf{b}} &= \left(\mathbf{M}'\mathbf{M}\right)^{-1}\mathbf{M}'\mathbf{Y} \\ &= \left(\mathbf{M}'\mathbf{M}\right)^{-1}\mathbf{M}'\mathbf{M}\mathbf{b} + \left(\mathbf{M}'\mathbf{M}\right)^{-1}\mathbf{M}'\mathbf{X}\mathbf{c} + \left(\mathbf{M}'\mathbf{M}\right)^{-1}\mathbf{M}'\boldsymbol{\varepsilon} \\ E(\hat{\mathbf{b}} \mid \mathbf{M}, \mathbf{X}) &= \mathbf{b} + \left(\mathbf{M}'\mathbf{M}\right)^{-1}\mathbf{M}'\mathbf{X}\mathbf{c}. \end{aligned} \tag{6}$$

The estimate of **b** is biased by $\left(\mathbf{M}'\mathbf{M}\right)^{-1}\mathbf{M}'\mathbf{X}\mathbf{c}$. This is "omitted variable" bias.

Clearly linear regression is useful because it can reduce bias caused by confounding variables. For example, students who do well on the PSAT (a pre-test for the SAT) may be less likely to get coached, but more likely to do well on the SAT. If this is the case, omitting PSAT scores as a covariate in the regression equation will result in a biased coaching effect estimate. A key point is that omitted variable bias is not the same thing as "selection bias." Selection bias occurs when the variable $COACH_i$ is endogenous—correlated to a latent covariate that has not been measured. If this is the case, the linear regression model generally will not produce unbiased estimates of the coaching effect—even if all the relevant observed covariates are included. The so-called Heckman Model (Heckman, 1978; 1979; Heckman & Robb, 1986; Greene, 1993), named after economist James Heckman who first developed the approach, has been applied in certain contexts as a general strategy for estimating a causal parameter in the presence of selection bias.

The Heckman Model

Under the Heckman Model, the variables in the regression equation are allowed to be correlated with the error term ε_i. In other words, the variables may be endogenous, so any causal parameter will suffer from selection bias.[6]

The motivation for the Heckman approach is essentially the same behavioral model as the one behind the use of linear regression:

$$f_i(COACH) = a + bCOACH + \mathbf{X}_i\mathbf{c} + \sigma\varepsilon_i \tag{7}$$

$$COACH_i = 1 \Leftrightarrow \alpha + \mathbf{X}_i\gamma + \delta_i > 0. \tag{8}$$

Everything in the causal relationship is the same as the one specified using the response schedule and selection function in (1) and (2). Observed SAT scores are again generated as

$$Y_i = f_i(COACH_i) = a + bCOACH_i + \mathbf{X}_i\mathbf{c} + \sigma\varepsilon_i, \tag{9}$$

where $COACH_i$ is determined by Equation 8. Assumptions i and ii are also retained:

i) $(\varepsilon_i, \delta_i)$ are iid in i with a standard normal distribution;

ii) $\{\mathbf{X}_i: i = 1, \ldots, N\}$ is independent of $\{\varepsilon_i, \delta_i: i = 1, \ldots, N\}$.

What has changed in the behavioral model? The critical change is that assumption iii is dropped. It is relaxed to allow ε_i and δ_i to be correlated. This introduces a new parameter, ρ, into the model. Under assumption iii of the linear regression model, the correlation ρ

[6] In this context, the term "selection bias" is being used synonymously with the term "endogeneity bias."

between ε_i and δ_i was restricted to 0. For the Heckman Model, ρ is allowed to take on any value between -1 and 1.

The causal parameter of interest is still *b*. Note that if ε_i and δ_i were not correlated, e.g. $\rho = 0$, then there would be no selection bias problem—linear regression could be used to correct for confounding and estimate an unbiased coaching effect. Intuitively, $\rho \neq 0$ will be the case if an unobserved reason why students decide to get coached is correlated with an unobserved reason that students perform well on the SAT. For example, suppose students with more "grit" are the ones most likely to get coached. At the same time, suppose students with more "moxie" will perform better on the SAT. (I offer no definition of grit and moxie; the two are distinguishable but latent.) While the linear regression model would assume that grit (i.e. δ_i) and moxie (i.e. ε_i) are independent, the Heckman Model allows for the possibility that they are correlated.

Given Equations 7-8 and assumptions i and ii, if $\rho \neq 0$ and the parameters *a*, *b* and **c** were estimated by regressing Y_i on a constant, $COACH_i$ and $\mathbf{X}_i$, the estimates would be biased. Because $\rho \neq 0$, the variable $COACH_i$ is endogenous, and $E(\varepsilon_i \mid COACH_i, \mathbf{X}_i) \neq 0$. The Heckman Model strategy is to get an estimate for this term, and then treat it as an observable confounder. Let $\lambda_i = E(\varepsilon_i \mid COACH_i, \mathbf{X}_i)$. If this value were known for student *i*, then regressing Y_i on a constant, $COACH_i$, $\mathbf{X}_i$ and λ_i would produce unbiased parameter estimates for *a*, *b*, **c** and *h*, where *h* is the regression coefficient associated with λ_i. Now, $E(\varepsilon_i - \lambda_i \mid COACH_i, \mathbf{X}_i) = 0$. If the assumptions of the Heckman Model are to be believed, then selection bias has been purged from the estimate of *b*.

In practice, λ_i is not known, but given the assumption that ε_i and δ_i have standard normal distributions, $\hat{\lambda}_i$ can be calculated as a function of the estimated parameters $\hat{\alpha}$ and $\hat{\gamma}$ in the selection function (8). Now, assuming that all confounding in the relationship between Y_i and $COACH_i$ is due to $\mathbf{X}_i$, and all selection bias is due to $\hat{\lambda}$, then regressing Y_i on a constant, $COACH_i$, $\mathbf{X}_i$ and $\hat{\lambda}$ will almost control for bias in the estimate of b due to both confounding and self-selection. Heckman (1979) has shown that $\hat{b}$ will converge to b asymptotically, so $\hat{b}$ will be biased but consistent. The details of the Heckman Model for the coaching application are sketched out below.

The starting point for the Heckman Model is the selection function describing the way students decide whether or not they will seek coaching. The vector $\mathbf{X}_i$ contains observable covariates related to the probability that a student is coached.[7] Latent covariates enter the picture through δ_i. The term δ_i is cast as an unmeasured latent continuous random variable with an assumed standard normal distribution. Student *i*'s decision to seek coaching is determined by a linear combination of the measured and unmeasured covariates represented by $\mathbf{X}_i$ and δ_i. The selection function specifies that if $\alpha + \mathbf{X}_i\gamma + \delta_i > 0$, student *i* will be coached. Otherwise, student *i* will not be coached. Given assumptions i and ii, another way of writing the selection function is

[7] In this setup, for the sake of parsimony, the covariates represented in $\mathbf{X}_i$ are the same in both Equation 7 and 8. This is not a restriction of the Heckman Model. It is possible for the covariates in the selection function to contain unique covariates related to the probability a student is coached, but not to subsequent SAT performance. Later I relax this notational restriction.

$$\begin{aligned} P(COACH_i = 1 \mid \mathbf{X}_i) &= P(\alpha + \mathbf{X}_i\gamma + \delta_i > 0 \mid \mathbf{X}_i) \\ &= P(-\delta_i < \alpha + \mathbf{X}_i\gamma \mid \mathbf{X}_i) \\ &= \Phi(\alpha + \mathbf{X}_i\gamma), \end{aligned} \tag{10}$$

where Φ represents the standard normal cumulative distribution function. Given all the $\mathbf{X}_i$'s, the $COACH_i$'s are assumed to be independent, so Equation 10 constitutes what is known as the probit model.

The following theorem[8] helps explain how the Heckman Model goes from specifying a selection function to getting an estimate for the bias term, $E(\varepsilon_i \mid \mathbf{X}_i, COACH_i)$.

Theorem I

Let t represent the point in the distribution at which a continuous random variable $v \sim N(0, 1)$ is truncated. When the truncation is from below

$$E(v \mid v > t) = \lambda(t) \tag{11}$$

$$Var(v \mid v > t) = 1 - \lambda(t)\left[\lambda(t) - t\right], \tag{12}$$

where

$$\lambda(t) = \frac{\phi(t)}{1 - \Phi(t)} \tag{13}$$

$$\phi(t) = \frac{1}{\sqrt{2\pi}} e^{\frac{-t^2}{2}} \tag{14}$$

$$\Phi(t) = \int_{-\infty}^{t} \phi(z)\,dz. \tag{15}$$

[8] For a proof of a more general version of this theorem, see Johnson & Kotz, 1970, 112-113. For a description consistent with the Heckman Model, see Greene, 1990, 682-689.

$\lambda(t)$ is commonly referred to as the Inverse Mills Ratio or Hazard Function. It is the ratio of the standard normal density function (14) to the normal cumulative distribution function (15). When the truncation in v is from above, then by symmetry of the normal distribution,

$$E(\nu \mid \nu \leq t) = \lambda(t) = -\frac{\phi(t)}{\Phi(t)}. \tag{16}$$

The goal is to estimate a value for the bias term $E(\varepsilon_i \mid \mathbf{X}_i, COACH_i)$ for student i. Fix a value $\mathbf{x}_i$ for $\mathbf{X}_i$. The selection bias term can be decomposed into two parts $E(\varepsilon_i \mid \mathbf{X}_i = \mathbf{x}_i, COACH_i = 1)$ and $E(\varepsilon_i \mid \mathbf{X}_i = \mathbf{x}_i, COACH_i = 0)$. Given the underlying behavioral model (Equations 7 and 8), and the condition that $COACH_i = 1$, it follows that δ_i no longer has a normal distribution, but a truncated normal distribution. Theorem I is used to compute the conditional expectation of δ_i, which will be $E(\delta_i \mid \alpha + \mathbf{X}_i\boldsymbol{\gamma} + \delta_i > 0)$. Similarly, under the condition that $COACH_i = 0$, it follows that δ_i again has a conditionally truncated distribution—this time the truncation is from above. Now the conditional expectation of δ_i is $E(\delta_i \mid \alpha + \mathbf{X}_i\boldsymbol{\gamma} + \delta_i \leq 0)$. The next step is to compute the conditional expectation of ε_i, given $\mathbf{X}_i$ and $COACH_i$.

Under the Heckman Model, ε_i and δ_i have correlation ρ. Let ξ_i be a random variable equal to $(\varepsilon_i - \rho\delta_i)/\sqrt{1-\rho^2}$. It follows from this definition that ξ_i has an expected value of 0 and is independent of δ_i. Think of ξ_i as the random variable that picks up the variance left unexplained if ε_i is regressed on δ_i. Now ε_i can be related to δ_i and ξ_i:

$$\varepsilon_i = \rho\delta_i + \sqrt{1-\rho^2}\xi_i. \tag{17}$$

Let $s_i = \alpha + \mathbf{X}_i\boldsymbol{\gamma}$. It follows from Equations 17 and 8 that

$$\begin{aligned} E(\varepsilon_i \mid \mathbf{X}_i = \mathbf{x}_i, COACH_i = 1) &= E(\varepsilon_i \mid \mathbf{X}_i = \mathbf{x}_i, s_i + \delta_i > 0) \\ &= \rho E(\delta_i \mid s_i + \delta_i > 0) \\ &= \rho E(\delta_i \mid \delta_i > -s_i). \end{aligned} \tag{18}$$

Note that ξ_i drops out of the equation because its conditional expectation is 0 by definition. The task is to evaluate the conditional expectation on the right side of (18). Taking advantage of the symmetry of the normal distribution and applying Theorem I leads to the Inverse Mills Ratio,

$$E(\delta_i \mid \delta_i > -s_i) = \frac{\phi(s_i)}{1 - \Phi(s_i)}. \tag{19}$$

Likewise,

$$\begin{aligned} E(\varepsilon_i \mid \mathbf{X}_i = \mathbf{x}_i, COACH_i = 0) &= E(\varepsilon_i \mid \mathbf{X}_i = \mathbf{x}_i, s_i + \delta_i \leq 0) \\ &= \rho E(\delta_i \mid s_i + \delta_i \leq 0) \\ &= \rho E(\delta_i \mid \delta_i \leq -s_i). \end{aligned} \tag{20}$$

This again yields the Inverse Mills Ratio

$$E(\delta_i \mid \delta_i \leq -s_i) = -\frac{\phi(s_i)}{\Phi(s_i)}. \tag{21}$$

It follows from (18-21) that

$$E(\varepsilon_i \mid \mathbf{X}_i, COACH_i) = \rho\lambda_i(COACH_i, s_i), \tag{22}$$

where

$$\lambda_i(COACH_i, s_i) = COACH_i\left(\frac{\phi(s_i)}{1 - \Phi(s_i)}\right) + (1 - COACH_i)\frac{-\phi(s_i)}{\Phi(s_i)}. \tag{23}$$

$\lambda_i(COACH_i, s_i)$ is a specific value for student *i*. While $\lambda_i(COACH_i, s_i)$ is not directly observable, it is estimable given the assumptions of the Heckman Model.

$\lambda_i(COACH_i, \hat{s}_i)$ is computed using (19), (21) and (23) after estimating parameter values for α and γ in (10) via maximum likelihood.

The behavioral model of (7) and (8) leads to

$$Y_i = a + bCOACH_i + \mathbf{X}_i\mathbf{c} + h\lambda_i(COACH_i, \hat{s}_i) + \varepsilon_i^* \quad (24)$$

where $\varepsilon_i^* = \sigma\varepsilon_i - h\lambda_i(COACH_i, \hat{s}_i)$. The causal parameter of interest is still *b*. The parameter *h* associated with $\lambda_i(COACH_i, \hat{s}_i)$ in Equation 24 is equal to $\sigma\rho$. Consistent estimates for *b* and *h* will be obtained by regressing Y_i on a constant, $COACH_i$, $\mathbf{X}_i$ and $\lambda_i(COACH_i, \hat{s}_i)$. Note that while it is $\hat{\sigma}\hat{\rho}$ that is estimated by $\hat{h}$, if an estimate for $\hat{\rho}$ is desired, it can be obtained by dividing $\hat{h}$ by $\hat{\sigma}$, where $\hat{\sigma}$ is estimated as a function of residuals from the regression equation. Because the conditional variance of ε_i^* depends on $\mathbf{X}_i$, a regression fit by OLS will be heteroskedastic. Estimates for *a*, *b*, **c** and *h* will be consistent, but inefficient. The standard errors estimated using OLS will be incorrect. A regression fit by Generalized Least Squares (GLS) will solve the latter problem (Greene, 1981). If the GLS estimate for *h* is statistically significant, this suggests that had *b* been estimated directly using linear regression without the Heckman correction, the estimate would have contained selection bias.

Finally, note that $\lambda_i(COACH_i, \hat{s}_i)$ essentially adds an interaction term consisting of $COACH_i$ and the Inverse Mills Ratio to the main effect for $COACH_i$ in the regression equation. The difference in expected SAT scores between coached and uncoached

students will be $\hat{b} + \hat{h}\left[\frac{\phi(\hat{\alpha}_i + \mathbf{X}_i\hat{\gamma})}{\Phi(\hat{\alpha}_i + \mathbf{X}_i\hat{\gamma})(1-\Phi(\hat{\alpha}_i + \mathbf{X}_i\hat{\gamma}))}\right]$. The effect of coaching estimated under the linear regression model is the combination of these two terms: the main coaching effect and the coaching by Inverse Mills Ratio interaction. The term in brackets will always be positive. The estimate $\hat{h}$ has been defined as the product of $\hat{\sigma}$ and $\hat{\rho}$. Since $\hat{\sigma}$ is always positive, if $\hat{\rho}$ is positive, this suggests that the coaching effect estimate from the linear regression model would be biased upwards. If $\hat{\rho}$ is negative, it suggests that the coaching effect estimate from the linear regression model would be biased downwards.

To summarize, the Heckman Model as applied to coaching studies has two main steps.

1. Specify a selection function for coaching status and estimate the parameters using maximum likelihood. Use these estimated parameters, and the assumed normal distributions of the response schedule and the selection function to compute the Inverse Mills Ratio when $COACH_i = 1$ and when $COACH_i = 0$.
2. Include $\lambda_i(COACH_i, \hat{s}_i)$ in a linear regression equation as a covariate. Estimate the coaching effect, $\hat{b}$ and the selection bias parameter, $\hat{h}$ (i.e. $\hat{\sigma}\hat{\rho}$) using OLS or GLS.

From Linear Regression to the Heckman Model

When a causal effect is estimated in an observational study, its interpretation is always threatened by the possibility of bias. Linear regression operates under the principal assumption that bias occurs because confounding variables were omitted from the regression equation. The Heckman Model assumes that bias comes from confounding caused by omitted variables, and more specifically, from endogeneity caused by the self-selection of subjects into treatment conditions. As presented here, the Heckman Model can be viewed as a two-step "correction" to the linear regression model in the presence of selection bias.[9]

Both linear regression and the Heckman Model assume that the functional form of the causal relationship between outcome, treatment and covariates is linear. In the context of observational studies where the coaching variable is dichotomous, the linearity assumption is violated if some or all of the covariates in $\mathbf{X}_i$ have a nonlinear relationship with Y_i. If the linearity assumption is incorrect, a coaching effect will be estimated as the difference between the wrong two regression surfaces. Both statistical approaches also typically make a constancy constraint, i.e. $b_i = b$, stipulating that person i = 1, ... , N is affected by the treatment in the same way. The constancy constraint is violated, for example, when certain types of students benefit significantly more or less from coaching. Indeed, interaction effects between coaching and student characteristics have been

[9] The Heckman Model can also be implemented as a one-step approach when estimation is done by maximum likelihood, but the two-step approach is more common in the applied literature (Vella, 1998).

analyzed from the very earliest coaching study by Dyer (1953) to the more recent study by Briggs (2001). If the constancy constraint is wrong, then causal inferences about "the" coaching effect may be misleading. Parametric assumptions such as linearity and constancy have been discussed in more detail in the context of an alternative approach to causal inference in observational settings known as the Propensity Matching Model. For details, see Rosenbaum & Rubin, 1983; 1984 and Rosenbaum, 2002.

A key difference between the two approaches is the relaxation of the independence assumption between ε_i and δ_i when going from linear regression to the Heckman Model. Normality was assumed for ε_i and δ_i throughout in order to focus attention on this difference. If normality does not hold, then the Heckman Model as described here falls apart as a correction for the selection bias problem. Normality is a necessary condition for consistent estimation under the Heckman Model, but not for linear regression. If the ε_i are iid, ε_i and δ_i are independent within student i, confounding covariates are included in the model, and the functional form is in fact linear, then linear regression will produce unbiased causal effect estimates even when the distribution of ε_i is non-normal.

Of course, the linear regression model can also serve descriptive or predictive purposes, with the well-known disclaimer that association does not imply causation. I have presented the rather strong assumptions necessary before association does imply causation. A clear problem in observational settings is that it is almost never realistic to assume that the bias in causal effect estimates is due solely to confounding from

measured covariates available to the investigating researcher. Generally speaking, the use of linear regression with covariates will at best only reduce omitted variable bias, not control or correct for it unequivocally.

Unlike linear regression, the Heckman Model is an approach specifically developed in the attempt to make unbiased causal inferences in observational settings. Because of the strong assumptions that underlie the model, its usefulness has been questioned by some statisticians (Wainer, 1986) and econometricians (Goldberger, 1983; Little, 1985). In one unusual case (Lalonde, 1986), the causal estimates from a Heckman Model were put to the empirical test—and the results were not encouraging. Lalonde gained access to data from a federally randomized experiment conducted to determine the average effect of a job training program. The effect was estimated by comparing the post-treatment incomes of subjects in an experimental treatment group to the post-treatment incomes of an experimental control group. Based on the findings from the randomized experiment, the average effect of the program appeared to be a little over $800, with a standard error of about $300. Lalonde attempted to recreate these results by substituting non-experimental control groups for the experimental control, and using a Heckman Model with different specifications of the selection function to approximate the result of the randomized experiment. The results showed that when using four different selection function specifications while holding constant gender and type of non-experimental control group, the estimated effect of the program varied from $10 to $670, and in few cases was the estimated effect within a standard error of the experimental estimate. Lalonde did not however, conclude that the Heckman Model's apparent sensitivity to

alternate selection function specifications threatened the usefulness of the model, nor did he speculate as to what drove this sensitivity.

Powers & Rock (1999) employed both linear regression and the Heckman Model to estimate a causal effect for SAT coaching in an observational setting. The findings from this study were that the two approaches produced relatively similar estimates of coaching effects, and that neither approach produced effect estimates considerably different from a baseline comparison with only pre-treatment test scores as covariates. In a footnote Powers & Rock reported that their Heckman Model estimates had been sensitive to specifications of the selection function, but no details were provided.

The relationship between the specification of the selection function and subsequent effect estimates would seem to merit closer attention, because as a procedure, the Heckman Model offers no guidance as to the covariates that should be included in its selection function. It is only assumed that $\{\mathbf{X}_i: i = 1, \ldots, n\}$ is independent of $\{\delta_i: i = 1, \ldots, n\}$. As a matter of identifiability, it does not matter whether the covariates in the selection function are different from those in the response schedule. The Inverse Mills Ratio is identified through its nonlinear relationship to $\mathbf{X}_i$. In some illustrations of the Heckman Model, it has been suggested that the covariates in the selection function should contain one or more variables related to the probability of treatment selection, but excluded from outcome prediction (e.g. Lalonde, 1986; Greene, 1993). In other illustrations, only covariates excluded from outcome prediction have been included in the selection function (e.g. STATA, 2000). Ideally, it would seem the choice of covariates

should be based on some theoretical understanding of the selection mechanism. I return to this issue in my empirical analysis of SAT coaching effects using the Heckman Model.

The NELS Data

The National Education Longitudinal Study of 1988 (NELS:88, hereafter referred to as "NELS") tracks a nationally representative sample of American students from the 8th grade through high school and beyond. The NELS data can be used for an observational evaluation of coaching effectiveness because it contains SAT scores and information about how students prepared for the SAT. A panel of nearly 15,000 students completed survey questionnaires in the second two waves of NELS in 1990 and 1992. One of these questions asked students to select from a range of options describing how they had prepared to take the SAT. In addition to student questionnaire responses, high school transcripts were collected. Each transcript included information on student grades, course taking patterns, school demographics, and college admission test scores.

For the analysis that follows, attention is focused on the NELS panel sample of students who completed surveys in the first (F1) and second (F2) follow-ups, and for whom transcript data was collected. This comprises an F1-F2 panel of 14,617 students. (For more information on the NELS sampling design, see the NELS Second Follow-up Student Component Data File User's Manual, 1995.) The emphasis in most SAT coaching studies has been on students who have taken the SAT and for whom there is a prior SAT or PSAT score available before a test preparation treatment has been

introduced. I similarly restrict attention to the 3,504 students from the NELS subsample who took both the PSAT and SAT, were members of the 10th grade and 12th grade cohorts as of the NELS F1 and F2 surveys, and indicated whether or not they had been coached as a means of preparing for the SAT.

The NELS Variables

To estimate a coaching effect from the NELS data using either linear regression or the Heckman Model requires three types of variables: an outcome variable (Y), a coaching variable ($COACH$) and covariates ($\mathbf{X}$). I briefly describe each in turn.

Math and Verbal SAT Scores

The outcome variable of interest is a score on either the math or verbal section of the SAT. As of the early 1990's, the SAT was a timed multiple choice test lasting for a total of two and a half hours. The test was then, and is now, intended to measure the constructs of mathematical and verbal reasoning, with scores from two different test sections. Each score was based on student responses to about 85 verbal items and 60 math items on the SAT. Because this is a relatively large number of items, and the items are chosen with great care, the SAT has the desirable technical feature of high internal consistency. The reliability of SAT math and verbal scores using Cronbach's Alpha is about .9, and the standard error of measurement for each test section is usually about 30 points. The mean and standard deviation of SAT-V scores (446 and 102) for the NELS

subsample are both slightly lower than the mean and standard deviation of SAT-M scores (501 and 117).[8] The mean scores for all college-bound seniors taking the test in 1991-92 was about 423 on the SAT-V, and 475 on the SAT-M. The mean SAT scores for the NELS subsample are slightly higher than those of the national population of test-takers because they are restricted to those students who had previously taken the PSAT.

The Coaching Variable

The treatment variable of interest is whether or not students have been coached before taking the SAT. The NELS F2 questionnaire asked students a targeted question about their test preparation activities. This question is replicated verbatim below.

To prepare for the SAT and/or ACT, did you do any of the following?	
A	Take a special course at your high school
B	Take a course offered by a commercial test preparation service
C	Receive private one-to-one tutoring
D	Study from test preparation books
E	Use a test preparation video tape
F	Use a test preparation computer program

With the exception of studying with a book, all of the methods listed above to prepare for the SAT have been classified as coaching in previous studies. In this analysis, students are classified as having been coached if they have enrolled in a commercial test preparation course. For a student answering question B above with a "yes", the dummy variable *COACH* is coded with a 1. For students answering with a "no", *COACH* is coded

[8] The SAT score scale was recentered as of 1995 (see Dorans, 2002 for details). Historical tables with mean SAT scores are now expressed in this metric. The mean scores for the NELS POP1 subsample correspond to recentered scores of 543 on the SAT-V and 524 on the SAT-M.

with a 0. The distinction made here is whether a test-taker has received systematic instruction over a short period of time. Preparation with books, videos and computers are excluded from the coaching definition because while the instruction may be systematic, it has no time constraint. Preparation with a tutor is excluded because while it may have a time constraint, it is difficult to tell if the instruction has been systematic. This definition of the term is consistent with that used by Powers & Rock (1999), and this makes the coaching effect estimates generated from the NELS data somewhat more comparable those generated from the nationally representative data in the Powers & Rock study. Also, commercial coaching is the most controversial means of test preparation, because it is costly, widely available, and comes with published claims as to its efficacy. About 15% of the students in the NELS subsample indicated that they had taken a commercial course to prepare for the SAT.

Covariates

To control for confounding in the estimation of coaching effects, an appropriate set of covariates must be chosen for $\mathbf{X}_i$. The choice of covariates can be guided to a great extent by previous investigations of coaching effectiveness. A review of the research literature on SAT coaching (see Briggs, 2002) indicates that previous SAT or PSAT scores, demographic characteristics, academic background and student motivation may serve to confound coaching effect estimates. Student motivation can be further divided into variables that proxy for intrinsic motivation (e.g. self-esteem) and extrinsic motivation (e.g. parental pressure). The latter variables may predict whether students are

likely to be coached, but are unlikely to have a direct influence on how students will perform on the SAT. Variables measuring extrinsic motivation should be particularly attractive candidates to include in a selection function for coaching as part of the Heckman Model.

When coached and uncoached students are compared along these sets of covariates in the NELS data, it appears that the coached group is more socioeconomically advantaged and more extrinsically motivated to take the SAT then uncoached counterparts. It is not clear that the coached group is necessarily comprised of academically "smarter" or more intrinsically motivated students—both groups are enrolled in college-preparatory classes, both performed about the same on NELS standardized tests in reading and math, both report having comparable levels of self-esteem, and both report that they do about the same amount of homework per week.

Analysis

Coaching effects can be estimated from the NELS data using both the linear regression model and Heckman Model. Earlier I described a behavioral model for SAT performance under which the coaching parameter *b* has a causal interpretation. This model is revisited with a slight modification below.

$$f_i(COACH) = a + bCOACH + \mathbf{X}_i\mathbf{c} + \sigma\varepsilon_i \tag{25}$$

$$COACH_i = 1 \Leftrightarrow \alpha + \mathbf{Z}_i\boldsymbol{\gamma} + \delta_i > 0\,. \tag{26}$$

$$Y_i = f_i(COACH_i) = a + bCOACH_i + \mathbf{X}_i\mathbf{c} + \sigma\varepsilon_i\,. \tag{27}$$

The selection function (26) has now been modified so that the covariates in the selection function ($\mathbf{Z}_i$) are allowed to be different from those in the response schedule ($\mathbf{X}_i$). This behavioral model forms the basis for any coaching effect estimated using linear regression or the Heckman approach.

Coaching effect estimates generated from linear regression or the Heckman Model cannot be compared directly to one another because they rely on different assumptions about the data structure, but they can be compared to the simplest alternative: the average SAT section score for coached students minus the average SAT score for uncoached students. For the SAT-V, this difference is 20 points (463 – 443); for the SAT-M, the difference is 30 points (526 – 496). If coached and uncoached students had been assigned randomly, these would be unbiased estimates of the coaching effects, and the usual method of determining the statistical significance of these differences could be used. Of course, the students in NELS were not randomly assigned, so these estimates are almost surely biased to some degree. What do linear regression and the Heckman Model suggest about the magnitude of this bias?

Coaching Effects and the Linear Regression Model

I start by specifying all covariates with a theoretical relationship to coaching status and SAT performance in the linear regression model. There are a total of 21 covariates in the linear regression model.

- Pre-coaching SAT scores: (*PSAT-V* and *PSAT-M*).
- Demographic characteristics: student age in years (*AGE*), socioeconomic status (*SES*)[10], dummy variables for gender (*FEMALE*), race/ethnicity (*ASIAN*, *BLACK*, *HISPANIC*, *AM_INDIAN*, *WHITE*), and whether the student's high school was public or private (*PRIVATE*), or located in a suburban, rural or urban locations (*SCH_URB*, *SCH_RUR*, *SCH_SUB*).
- Academic background: dummy variables for whether or not a student reports having taken an Advanced Placement class (*AP*) or remedial classes in math (*RE_MATH*) or English (*RE_ENG*); a dummy variable indicating whether or not the student has been enrolled in a rigorous academic program while in high school (*RIGHSP*); scores on standardized achievement tests in math (*F1MATH*) and reading (*F1READ*) administered as parts of the NELS survey, the number of units a student has taken in college preparatory math courses[11] (*MTHCRD*), and his or her weighted grade point average in those courses (*MTHGRD*).
- Intrinsic student motivation: the NELS self-esteem (*F1ESTEEM*) and locus of control (*F1LOCUS*) indices, and a dummy variable indicating whether the student

[10] The SES index was developed as part of the NELS database, and combines information about parental education, income and occupation into a single variable. Generally, students with higher SES values come from families with parents that are better educated, wealthier and have jobs in more prestigious occupations. For the NELS subsample considered here, the SES index has a mean of .44, a standard deviation of .73, and a range from –2.4 to 2.5.

[11] College preparatory math courses consist of algebra, geometry, trigonometry, pre-calculus and calculus.

reported averaging more than 10 hours per week on homework during high school (*HOMEWORK*).

The reference categories are *WHITE* and *SCH_SUB* for the racial/ethnic and school location dummy variables respectively.

Table 1. Coaching Effects using the Linear Regression Model

	SAT-V (mean = 447, sd = 101)			SAT-M (mean = 504, sd = 116)		
R^2	.788			.822		
adj R^2	.787			.818		
Coached/Total	503/3144			503/3144		
Variables in Regression Eqn	$\hat{a},\hat{b},\hat{c}$	Std Error Range *DEFF* = 1	*DEFF* = 3	$\hat{a},\hat{b},\hat{c}$	Std Error Range *DEFF* = 1	*DEFF* = 3
Constant	144.1	36.1	63.6	-7.6	37.5	66.1
COACH	11.1*	2.4	4.3	19.2*	2.5	4.5
PSAT-M	.05*	.02	.03	.41*	.02	.03
PSAT-V	.61*	.01	.02	.09*	.01	.02
AGE	-8.7*	1.9	3.4	-2.7	2.0	3.5
SES	3.8	1.4	2.4	10.2*	1.4	2.5
FEMALE	-5.0	1.9	3.3	-16.1*	1.9	3.4
ASIAN	7.9	3.5	6.2	4.8	3.6	6.4
BLACK	-3.5	3.2	5.6	-14.3*	3.3	5.8
HISPANIC	-3.1	3.4	6.1	-4.6	3.6	6.3
AM_INDIAN	-6.2	14.4	25.4	-26.2	15.0	26.4
PRIVATE	8.9*	2.4	4.2	-0.9	2.5	4.4
SCH_RUR	-6.6	2.3	4.0	-3.5	2.4	4.1
SCH_URB	1.1	2.0	3.6	1.3	2.1	3.7
AP	12.4*	1.9	3.3	8.8*	2.0	3.5
RE_ENG	-11.4	4.2	7.4	8.2	4.4	7.7
RE_MATH	1.7	4.0	7.1	-19.1*	4.2	7.3
RIG_HSP	-1.2	1.7	3.1	2.8	1.8	3.2
F1READ	2.5*	0.2	0.3	-0.5	0.2	0.3
F1MATH	0.4	0.2	0.4	4.9*	0.2	0.4
MTHCRD	-1.3	1.3	2.3	8.8*	1.3	2.4
MTHGRD	3.6	1.4	2.4	14.8*	1.4	2.5
F1ESTEEM	5.2	1.6	2.8	-1.9	1.6	2.9
F1LOCUS	-6.2	1.8	3.2	-2.1	1.9	3.4
HOMEWORK	3.5	1.8	3.1	1.4	1.9	3.3

* p-value for two-sided t-test < .05 across SE range
DEFF = design effect correction

Table 1 reports the results of separate linear regressions of student SAT-V and SAT-M scores on a constant, *COACH*, and the full set of 21 covariates in $\mathbf{X}_i$ listed above. Each regression was weighted by the variable *DESWGT* to account for the NELS

population weights, as well as the design effects caused by the stratification and clustering of students in the NELS sample (see Appendix A for details). Regressions were run with two different versions of *DESWGT*; one with a design effect correction set equal to 1 (e.g no design effect), the other with a correction set equal to 3. The clustering of students in the POP1 subsample, amounts to a mean of 4 and median of 6 students per school—relative to a mean and median of 14 for the full F1-F2 panel sample. In the NELS subsample there is on average just one coached student per sampled school. Given this, using a design effect correction of 3 will probably overestimate standard errors. All else being equal, the standard errors of parameter estimates associated with each version of the *DESWGT* variable should reflect lower and upper bounds in tests of statistical significance, and to give a sense for this range, both are reported for the regression coefficient estimates in Table 1.

Under the linear regression model, the estimated effect for *COACH* is 11 and 19 points respectively on the SAT-V and SAT-M. Expressed as a proportion of a standard deviation in SAT scores, this amounts to effect sizes of .11 and .16 for each estimate. Both effects are statistically significant whether tested using the standard errors based on the lower or upper design effect bounds. Using the more conservative standard error estimate, the 95% confidence intervals for the estimated SAT-V and SAT-M coaching effects are [3, 20] and [10, 28]. These estimated effects suggest that the linear regression model reduces bias due to confounding. After including the covariates $\mathbf{X}_i$ in the model, the estimated SAT-V coaching effect decreases by 9 points from 20 to 11, and the estimated SAT-M coaching effect decreases by 11 points from 30 to 19.

On the whole, the estimated coaching effects and associations of covariates with SAT-V and SAT-M scores in the linear regression model seem reasonable. Still, the possibility that one or more is biased cannot be ruled out. One possible source of bias may be additional covariates that have been mistakenly omitted from the regression model. For example, perhaps the correct model would include a series of interaction terms with the coaching variable (c.f. Briggs, in press). Another possibility is that bias exists of a very specific nature due to the endogeneity of the variable *COACH*. This latter problem is one that the Heckman Model has been designed to solve.

Coaching Effects and the Heckman Model

Specifying a Selection Function

In order to estimate an effect for *COACH* using the Heckman Model, I start by specifying a selection function that, given a set of covariates $\mathbf{Z}_i$, predicts whether student i will be coached or not. The specification decision hinges upon what covariates are included in $\mathbf{Z}_i$. Ideally, students in the NELS survey would have been asked questions about why they did or did not enroll in coaching programs, but as NELS was not designed with the Heckman Model in mind, such data is not available. This is a fairly typical situation in an observational study. As a consequence, the specification of a selection function is seldom guided by theory. In many empirical applications of the

Heckman Model, the decision of what covariates to include in $\mathbf{Z}_i$ appears to be largely a matter of ensuring that the model is well identified.

Figure 3. Five Selection Function Specification

SF1	$\mathbf{Z}_i = \{\mathbf{X}_i\}$
SF2	$\mathbf{Z}_i = \{\mathbf{X}_i, PARENT_i\}$
SF3	$\mathbf{Z}_i = \{PARENT_i, PPRESS_i, HWTUTOR_i, HI\_MOT_i\}$
SF4	$\mathbf{Z}_i = \{SES_i, SCH\_RUR_i, REMATH_i, MTHCRD_i, PPRESS_i, HWTUTOR_i, HI\_MOT_i\}$
SF5	$_i = \{AGE_i, SES_i, SCH\_RUR_i, MTHGRD_i, PARENT_i, PPRESS_i, HWTUTOR_i, HI\_MOT_i\}$

SF2, SF3, SF4 and SF5. The predictors in each specification are listed in Figure 3. Which of these is the "right" specification of the selection function? A reasonable case could be made for each of the five. In SF1, all the covariates specified as possible confounders in the regression equation are included as predictors in the selection function, and this represents the kind of mechanical use of the Heckman Model to be expected when the data analyst has no operating theory for how students select themselves into coaching. Note that the Heckman Model in this case is identified only by the nonlinearity of the selection function. Some have referred to this as "weak" identification (Breen, 1996; Vella, 1998). In SF2, one additional predictor, the dummy variable *PARENT*—which takes a value of 1 if a student was strongly encouraged by his or her parents to prepare for the SAT—has been added to the selection function. Now the model is overidentified, since *PARENT* is not a covariate in the response schedule. Here we imagine the data analyst has access to at least one variable thought to predict coaching status, but not SAT performance. This is known as a single exclusion restriction. SF2 doesn't constitute a theory per se, but it is the simplest possible improvement over SF1. For SF3, only covariates excluded from $\mathbf{X}_i$ in the linear regression equation are included as predictors in

the selection function[12], where *PPRESS*, *HWTUTOR* and *HI_MOT* are dummy variables that take values of 1 if the student's test preparation plans were "often" discussed with his or her parents, if the student had a private tutor that helped with homework during high school, and if the student did poorly on the PSAT relative to his high school GPA in math courses. Under SF3, there are now four variables thought to predict coaching status, but not SAT performance. In addition, the strong and questionable assumption is made that no covariates in $\mathbf{X}_i$ should be used to predict coaching status. The specification SF3 is meant as an extreme contrast with SF1. In SF1, all covariates in $\mathbf{X}_i$ are also in $\mathbf{Z}_i$; in SF3, no covariates[13] in $\mathbf{X}_i$ are also in $\mathbf{Z}_i$. In SF4, all predictors included in the selection function are chosen by a stepwise selection algorithm. SF4 is another example of a mechanical approach a data analyst might take in specifying the selection function: all possible covariates are thrown into an algorithm, and an optimal subset emerges. Finally, for SF5, predictors are chosen for two reasons: because they have some theoretical relationship to coaching status (*SES*, *PARENT*, *PPRESS*, *HWTUTOR*, *HI_MOT*) or because they have an empirical relationship to coaching status (*AGE*, *SCH_RUR*, *MTHGRD*). SF5 is an approximation of a theory-based specification approach. Here the

[12] Values for the predictors *PARENT*, *PPRESS* and *HWTUTOR* were missing for anywhere from 2 to 10% of the NELS subsample of 3,144 students used in the linear regression model. To ensure that subsequent Heckman Model parameter estimates will be based on the same sample of students as those produced by linear regression, missing values for these predictors were coded as three unique dummy variables which took the value of 1 if a student's response was missing, and 0 otherwise. For any selection function specification including one or more of these three variables, the associated missing value dummy variable *MPARENT*, *MPPRESS* or *MHWTUTOR* was also included.

[13] Strictly speaking this is not true since *HI_MOT* is itself a function of *PSAT-V*, *PSAT-M* and *MTHGRD*.

data analyst has taken some care in choosing predictors with a hypothesized relationship to coaching status (i.e. it is well-established that coaching programs can be expensive, and hence high-SES students are more likely to enroll in them). In addition, the data analyst has analyzed the pairwise cross-tabulations of all covariates with coaching status, and included three for which there was evidence of a statistically significant relationship. SF5 has four exclusion restrictions as in SF3, but includes in $\mathbf{Z}_i$ a subset of covariates from $\mathbf{X}_i$, as in SF4.

Table 2 presents the parameter estimates generated from a weighted probit model (weighted by the variable $DESWGT_i$ with a design effect correction of 3) for each of the five SF specifications. It is not at all obvious on statistical grounds that any one of the five specifications is the best choice for use in the Heckman Model. Unlike linear regression, where model fit is often assessed on the basis of R^2, there is no such measure of absolute fit for the probit model. When compared using a likelihood ratio (LR) test to a baseline specification with just a constant and no predictors, all five SF specifications would be considered a statistical improvement. A variant of this approach is represented by the "Pseudo R^2" values in the third row of Table 2. The Pseudo R^2 for each specification is calculated as $(1 - L)/L_0$, where L is the log likelihood for a given specification of the selection function, and L_0 is the log likelihood for the baseline specification. According to this criterion, the SF4 and SF5 specifications improve model fit the best relative to the baseline model, but not by much—all five specifications are within about .04 of one another. Of the five specifications, only SF1 and SF2 are nested and can be compared directly using a likelihood ratio test. The difference in deviance

between SF2 and SF1 is 11.7 with an approximate Chi-Square distribution on 2 degrees of freedom. On this basis SF1 can be rejected in favor of SF2, but no LR test can recommend SF2 over SF3, SF4 or SF5.

Table 2. Selection Function Parameters Estimated using Weighted Probit Model

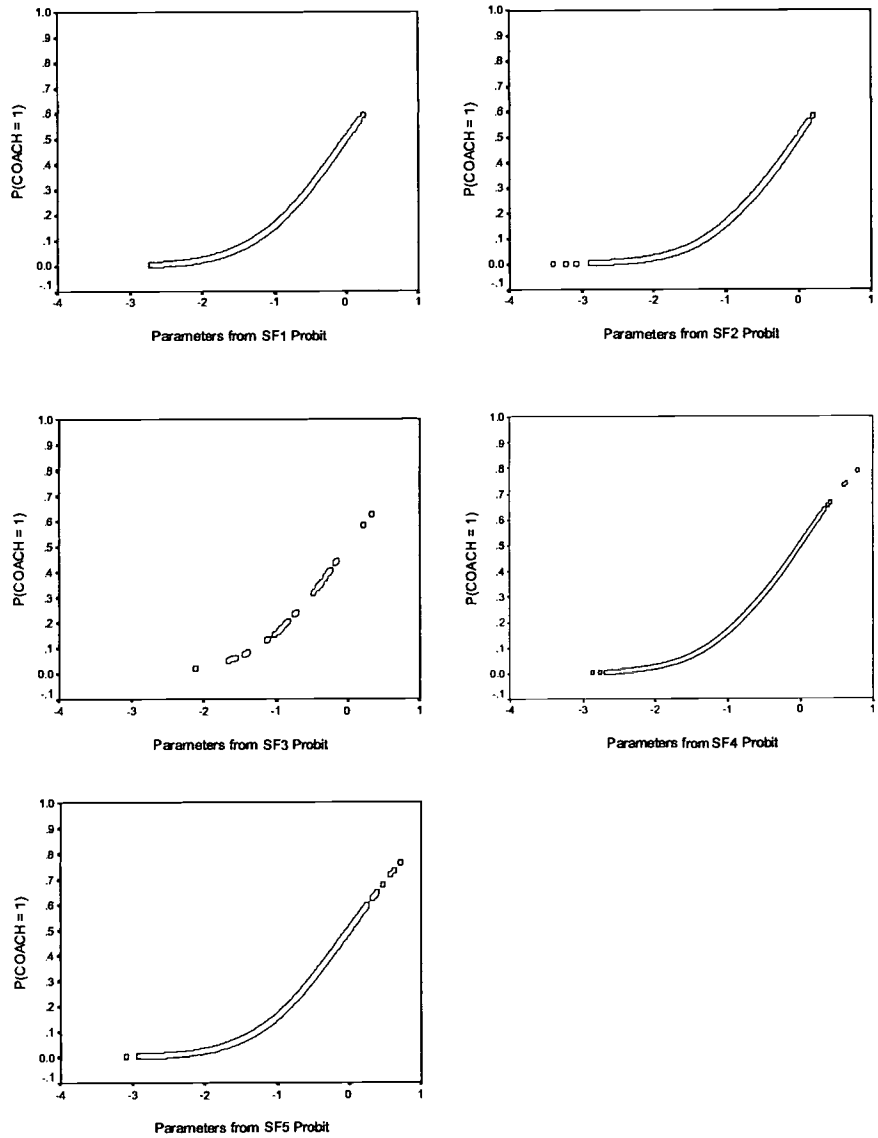
	SF1		**SF2**		**SF3**		**SF4**		**SF5**	
Log Likelihood	-1175.3		-1163.6		-1187.3		-1119.2		-1119.2	
dof	23		25		7		8		11	
Pseudo R^2	.0994		.1084		.0902		.1424		.1423	
% sig covariates	13% (3/23)		20% (5/25)		86% (6/7)		100% (8/8)		72% (8/11)	
Variables in Selection Fcn	$\hat{\alpha},\hat{\gamma}$	se	$\hat{\alpha},\hat{\gamma}$	se	$\hat{\alpha},\hat{\gamma}$	se	$\hat{\alpha},\hat{\gamma}$	se	$\hat{\alpha},\hat{\gamma}$	se
Constant	-3.984*	1.886	-4.712*	1.921	-2.115*	.187	-2.146*	.234	-4.202*	1.870
PSAT-M	-.0006	.0007	-.0006	.0007						
PSAT-V	-.0004	.0006	-.0003	.0006						
AGE	.142	.099	.142	.100					.112	.102
SES	.563*	.091	.548*	.091			.441*	.078	.439*	.079
FEMALE	.084	.096	.084	.096						
ASIAN	.128	.153	.138	.154						
BLACK	.078	.170	.097	.170						
HISPANIC	-.031	.163	-.028	.166						
NATIVE	-.326	.518	-.342	.518						
PRIVATE	.058	.146	.061	.148						
SCH_RUR	-.390*	.116	-.374*	.117			-.429*	.124	-.416*	.120
SCH_URB	.065	.159	.066	.159						
AP	-.052	.142	-.049	.143						
RE_ENG	.151	.200	.149	.199						
REMATH	.300	.199	.307	.194			.471*	.161		
RIG_HSP	.093	.108	.092	.108						
F1READ	.001	.008	.001	.008						
F1MATH	-.010	.009	-.010	.009						
MTHCRD	.143*	.058	.139*	.058			.138*	.055		
MTHGRD	.159	.113	.161	.113					.009	.057
F1ESTEEM	.114	.078	.117	.077						
F1LOCUS	-.093	.093	-.097	.093						
HOMEWORK	.006	.097	-.003	.097						
PARENT[a]			.695*	.191	.702*	.187			.602*	.188
MPARENT[a]			.745*	.220	.721*	.230			.688*	.231
PPRESS[a]					.677*	.130	.652*	.115	.628*	.115
MPPRESS[a]					.529*	.145	.552*	.149	.526*	.143
HWTUTOR[a]					.459*	.113	.333*	.121	.334*	.121
MHWTUTOR[a]					.560	.394			.592	.370
HI_MOT[a]					.472*	.233	.424*	.210	.447*	.205

* p-value for two-sided t-test < .05 (DEFF = 3)
N = 3,144
[a] These covariates are excluded from the regression equation

Another possible criterion to consider in picking a "best fitting" specification is one with the largest proportion of statistically significant probit coefficient estimates. This is fairly important, since the next step of the Heckman Model is to calculate an Inverse Mills Ratio as a function of the estimated coefficients, whether they are significant or not. Naturally, the SF4 specification comes out on top here—all of its coefficients are statistically significant, because its predictors were selected with this criterion in mind. The SF3 and SF5 specifications are not far behind, with 86% and 72% of estimated coefficients statistically significant. SF1 and SF2 are particularly weak relative to this criterion, with only 13% and 20% of estimated coefficients statistically significant.

For each of the $k = 1$ through 5 SF specifications, let $\hat{s}_{ik} = \hat{\alpha}_k + \hat{\gamma}_k \mathbf{Z}_i$. Figure 4 shows the plots of the predicted probabilities of being coached as a function of $\hat{s}_{ik}$. The shape of the five curves is generally quite similar, though for SF4 and SF5 the highest estimated probability is about .2 higher at the maximum value of $\hat{s}_{ik}$. In terms of the actual and predicted number of coached students for each specification, all the specifications tend to underpredict the number of coached students. None of these models predicts correctly the coaching status for more than about 20% of those students who were actually coached.

Figure 4. Predicted Probabilities of COACH = 1 for SF Specifications



The point of these model comparisons is that in most applications of the Heckman Model, precious little ink has been spent validating selection function specifications. Seldom are alternate specifications compared, and it is even more seldom that there is any theory to bolster the specification ultimately chosen. The decision of what predictors

to include or exclude from the selection function is a non-trivial one, and can have substantial ramifications on the estimated parameters generated by the Heckman Model.

Heckman Model Estimates

Using Equation 23, $\lambda_{ik}(COACH_i, \hat{s}_{ik})$ can be estimated for the $k = 1, ..., 5$ SF specifications. For the second step of the Heckman Model I proceed by including $\lambda_{ik}(COACH_i, \hat{s}_{ik})$ as a covariate in the regression of Y_i on a constant, $COACH_i$, and $\mathbf{X}_i$. The covariates in $\mathbf{X}_i$ are identical to those specified for the linear regression model. All cases are weighted by $DESWGT_i$ with a design effect correction of 3. In addition, because the conditional variance of ε_i under the Heckman Model is heteroskedastic, a generalized least squares fitting procedure (Greene, 1981) is used to get efficient standard error estimates for the regression coefficients. Table 3 reports the results of these regressions for SAT-V and SAT-M test scores.

Table 3. SAT Coaching Effects using the Heckman Model

	SAT-V			SAT-M		
	$COACH_i$	$\lambda_i(COACH_i, \hat{s}_i)$	$\hat{\rho}$ of $(\delta_i, \varepsilon_i)$	$COACH_i$	$\lambda_i(COACH_i, \hat{s}_i)$	$\hat{\rho}$ of $(\delta_i, \varepsilon_i)$
SF1	69* (30)	-32* (16)	-.60	79* (30)	-33* (17)	-.64
SF2	58* (26)	-26 (14)	-.42	59* (28)	-22 (15)	-.36
SF3	0 (15)	7 (8)	.15	30 (16)	-6 (9)	-.10
SF4	17 (15)	-3 (9)	-.05	46* (16)	-16 (9)	-.25
SF5	12 (15)	-1 (8)	-.01	42* (15)	-13 (9)	-.20

N = 3,144 [effective sample size after design effect correction = 1,015]
* p-value < .05 (based standard errors with design effect = 3)

SF1 = all covariates in regression eqn used in selection eqn
SF2 = all covariates in regression eqn + 1 covariate (*PARENT*) not used in reg eqn
SF3 = only covariates not used in reg eqn, all dummies (*HWTUTOR, PARENT, PPRESS, HI_MOT*)
SF4 = covariates chosen by stepwise selection (*SCH_RUR, PPRESS, HWTUTOR, REMATH, HI_MOT, SES, MTHCRD*
SF5 = covariates that were stat sig in coaching crosstabs (*AGE, SES, MTHGRD, SCH_RUR, HWTUTOR, PARENT, PPRESS, HI_MOT*)

The estimated effects for *COACH* vary, sometimes dramatically, depending upon which version of $\lambda_{ik}(COACH_i, \hat{s}_{ik})$ is included in the Heckman Model. For specifications with *SAT-V* as the dependent variable, the estimated coaching effect ranges from a low of 0 points to a high of 69 points. For specifications with *SAT-M* as the dependent variable, the estimated coaching effect ranges from a low of 30 points, to a high of 80 points. Parameter estimates for covariates under all five specifications of the Heckman Model with either *SAT-V* or *SAT-M* as the dependent variable were generally similar to those from the linear regression model.

Depending upon the selection function that is specified, the Heckman Model tells a different story about the nature of selection bias in SAT coaching. In models with SAT-V as the dependent variable, the estimated correlation $\hat{\rho}$ between δ_i and ε_i is -.60 and -.42 for SF1 and SF2, but close to zero for SF4 and SF5. When SAT-M is the dependent variable, the estimated correlation is -.64 for SF1, but between -.36 and -.10 for SF2 through SF5.

Only in the SF1 specification of the model is the parameter estimate for $\lambda_{ik}(COACH_i, \hat{s}_{ik})$ also statistically significant, indicating the presence of selection bias. For these (as well as most other) specifications, the estimated negative correlations between δ_i and ε_i would suggest that the students who are more likely to get coached are the ones who are *less* likely to perform well on a particular section of the SAT. If these versions of the Heckman Model are to be believed, it would indicate that the coaching

effects estimated by the linear regression model will be biased downwards. On the other hand, most specifications of the Heckman Model considered here suggest that any selection bias in the data is not statistically significant.

Multicollinearity helps explain why coaching effect estimates vary so dramatically, with large standard errors, under different specifications of the Heckman Model selection function. In particular, the variable $COACH_i$ and $\lambda_{ik}(COACH_i, \hat{s}_{ik})$ are strongly correlated, which follows from the fact that the latter is defined as an interaction with the former. When the $\lambda_{ik}(COACH_i, \hat{s}_{ik})$ based on SF1 and SF2 are regressed on a constant, $COACH_i$ and $\mathbf{X}_i$, the respective adjusted R^2's are .98 and .97. Likewise, the regressions based on SF3, SF4 and SF5 have adjusted R^2's of .92, .94 and .92.

To see more clearly the collinear relationship between the variable $COACH_i$ and $\lambda_{i5}(COACH_i, \hat{s}_{i5})$, I subtract from each variable its predicted value when regressed on $\mathbf{X}_i$. The resulting variable is the residual component not predicted by $\mathbf{X}_i$. The two residualized variables— $COACH_i r$ and $\lambda_{i5}(COACH_i, \hat{s}_{i5})r$ —are plotted in Figures 5 and 6 for the conditions $COACH_i = 1$ and $COACH_i = 0$. The correlation between the residualized variables is still about .73.

Figure 5. Collinearity when COACH = 1 (ρ = .72)

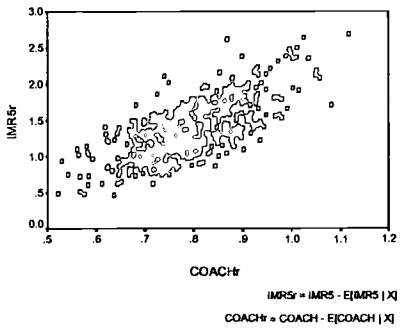
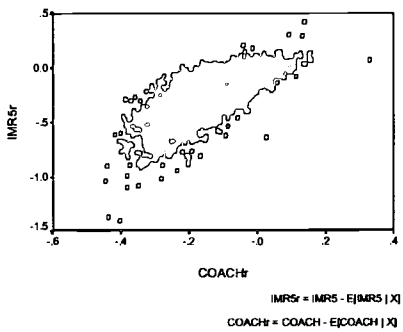


Figure 6. Collinearity when COACH = 0 (ρ = .74)



The easiest solution to the multicollinearity problem is to omit one or more covariates from the regression equation. But this is no real solution to the problem because the underlying behavioral model has now been violated—any decrease in multicollinearity will come with a potential increase in bias. Other solutions have been proposed and applied to handle collinear data without omitting variables (c.f. ridge regression and principal components analysis described in Greene, 1993, p. 270-273). A detailed discussion of these methods is outside the scope of this paper, but it is important to note that "solutions" to multicollinearity have their own associated problems. To the extent that such methods change the structure and relationship of the data under

consideration, they will almost certainly change the causal interpretation of the Heckman Model as presented here.

Empirical Comparisons

Figures 7 and 8 compare the estimated SAT-V and SAT-M coaching effects estimated by 1) taking the difference in average scores between coached and uncoached students, 2) using linear regression and 3) using the five Heckman Model specifications. I include around each point estimate the corresponding 95% confidence interval.

Figure 7. Comparison of SAT-V Coaching Effect Estimates

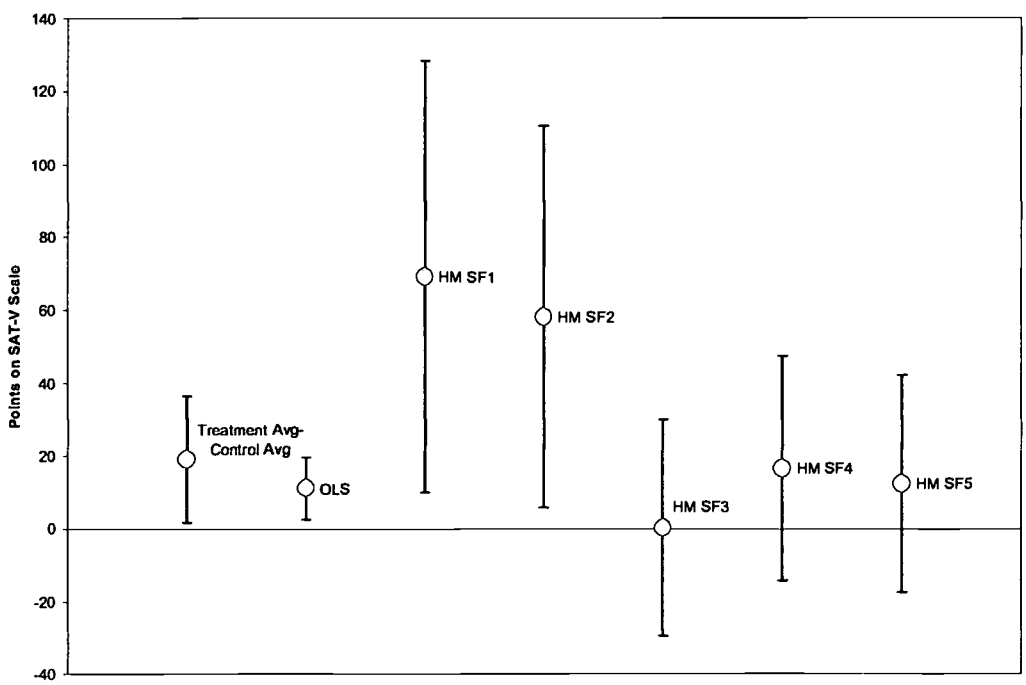
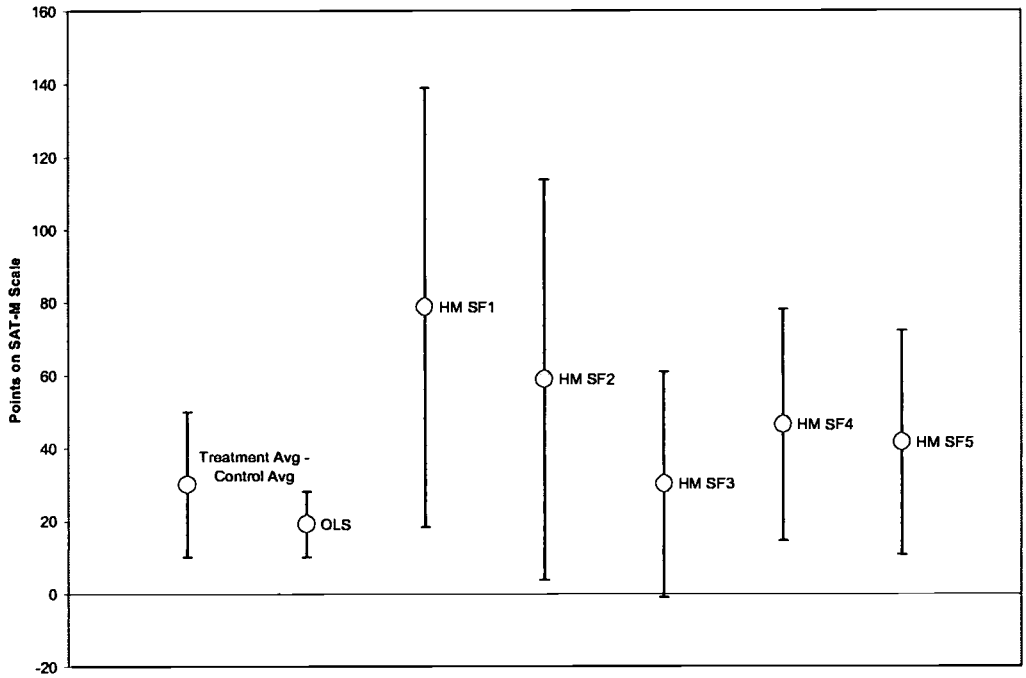


Figure 8. Comparison of SAT-M Coaching Effect Estimates



For the SAT-V, the linear regression model produces a statistically significant point estimates of about 11 points for the coaching effect. The Heckman Model produces effect estimates ranging from 0 to 70 points, only two of which (SF1 and SF2) are statistically significant. If the SF1 and SF2 specification of the Heckman Model are ignored, the SAT-V effect estimates from both models are smaller than what would be estimated by simply taking the average difference in SAT-V scores for coached and uncoached students. For the SAT-M, the Heckman Model produces coaching effect estimates ranging from 30 to 70 points—estimates that are generally more than twice as large as the 19 point estimate produced under linear regression. The SAT-M coaching effect estimates tend to be statistically significant under both models. Under the Heckman Model the estimates tend to be larger (SF 3 is the exception) than what would be

estimated by simply taking the difference in the average SAT-M scores for coached and uncoached students, while under linear regression the estimate is smaller.

Unlike the Lalonde study, there is no absolute criterion against which to compare the coaching effects estimated by the Heckman Model. Only the Powers & Rock study has used the Heckman Model to estimate coaching effects. The covariates and predictors available in the Powers & Rock data, while not quite of the same quality as some of those available from NELS, were fairly similar. In their regression equation Powers & Rock included covariates for PSAT or first SAT scores, father's education, student high school GPA, math GPA, race/ethnicity and two measures of student motivation.[14] Their selection function included all the same variables, and also included student's GPA in high school social science courses. This specification of the Heckman Model is probably most comparable to my SF2. Yet Powers & Rock's SAT-V coaching effect estimate (12 points) produced using the Heckman Model was similar only to those produced under SF4 and SF5 with the NELS data; for the SAT-M their effect estimate (13 points) was generally less than a third of the NELS-based estimates. Powers & Rock also estimated standard errors that were on the whole much smaller than those found in the analysis of the NELS data, in part perhaps because their data structure did not require a design effect correction.

[14] This information was not included in their published study of 1999, but was provided to me in a personal communication (Rock, 2002).

Discussion

This paper has hopefully shed some light on the use of the Heckman Model to estimate unbiased causal effects with observational data. Extreme caution should be exercised before applying the Heckman Model as a means of drawing causal inferences about a treatment effect. There is seldom any theory to guide the specification of the selection function, and if the selection function is specified just with the objective of identifying the model (e.g. SF1 and SF2), the resulting effect estimates will probably be highly questionable, if not completely out of whack. Once a selection function has been specified, estimated, and used to calculate the Inverse Mills Ratio, the next concern should be the potential for multicollinearity between the covariates, the treatment variable, and the interaction between the treatment variable and the Inverse Mills Ratio, with most of the problem stemming from the collinearity among the latter terms. When multicollinearity is a problem, it may cast doubt on both the estimated treatment effect and the standard errors around the treatment effect. All too often the Heckman Model has been applied in the social science with little to no discussion of these issues. With access to the right software (e.g. STATA, LIMDEP), the Heckman Model is easily implemented with seemingly obvious causal conclusions. I would suggest that when this is takes place without a compelling theoretical rationale and a careful scrutiny of the data, such conclusions are of dubious value.

In general, researchers must be quite cautious in using statistical models to draw causal conclusions, particularly given the types of assumptions that must be invoked.

There is no statistical silver bullet. In the social sciences, bias in the estimated effects from any given study is very difficult to rule out, no matter how intuitively appealing the methodology. A point worth emphasizing is that the best way to establish a causal effect from observational data, irrespective of the statistical model being used, is to replicate the results with a different sample. There was no single study or statistical model that established from observational data the deleterious effects of smoking on a range of health outcomes. Rather it was the consistent replication of these findings over a long period of time that led the way to what is now an accepted causal relationship. It is unfortunate that this approach has seemingly had limited traction in the educational research literature.

Appendix: Population Weights and Design Effects

Sample weights have been constructed and made available as part of the NELS database to allow for population inferences from the longitudinal and cross-sectional samples. To make the F1-F2 panel sample representative of the national population of 10^{th} to 12^{th} grade students during the 1990 to 1992 period, the NELS weight *F2TRP2WT* is applied to all statistical analyses that follow. Use of this weight indicates that the F1-F2 NELS panel is representative of an underlying population of about three million students. Since the NELS F1-F2 panel is generated from a stratified cluster sample (SCS), the estimated standard errors of population parameters (e.g. the mean for a particular transcript variable or survey item response) will generally be larger than the standard errors that would be estimated had the panel been generated from a simple random sample (SRS). The ratio of these two standard error estimates for any given parameter corresponding to the variable j is known as a *design effect* ($DEFF_j$). That is

$$DEFF_j = \frac{SE_j(SCS)}{SE_j(SRS)}.$$

The standard errors estimated by typical statistical software packages such as SPSS, STATA or SAS are generally calculated under the assumption that the data has come from a SRS. The larger the design effect, the more that standard errors erroneously calculated under an SRS assumption will underestimate the standard errors that befit the SCS sample design of NELS. Essentially, the clustering of the NELS sample decreases the effective sample size because students sampled within the same school are not statistically independent. Note that this violates a common assumption of both linear regression and the Heckman Model, namely, that ε_i and δ_i are each independently

distributed across students. If this lack of independence is not taken into account, tests of significance using estimated standard errors that are too small may well result in Type I errors.

A school identification code is available for 13,471 students (92%) in the NELS F1-F2 panel. These students were sampled from 974 different high schools. The mean and median size of the student clusters per school is 14. According to the NELS F2 manual this corresponds to a mean and median design effect across all variables of about 3.7 and 3. For subsamples of students in the F1-F2 panel, the mean and median cluster sizes, and presumably the corresponding design effects will be smaller. Finding out just how much smaller is outside the scope of this study. For the analyses that follow, all standard errors are estimated using proportional population weights that include a design effect correction to reduce the effective sample size. This amounts to a first order approximation of the standard errors that would be estimated under the assumption of a SCS.

More specifically, denote each student in the NELS F1-F2 panel sample with the subscript i. For any subset of S cases taken from the F1-F2 panel sample, the NELS variables that correspond to student i are weighted by the variable $DESWGT_i$, where

$$DESWGT_i = \frac{1}{DEFF}\left(\frac{F2TRP2WT_i}{\frac{1}{S}\sum_{i=1}^{S} F2TRP2WT_i}\right).$$

$F2TRP2WT_i$ is the population weight of cases in the F1-F2 panel sample for whom transcript data was collected, and $DEFF$ is a postulated design effect that applies

to all NELS variables. As an approximation of the design effect associated with each variable, it is assumed that $DEFF_j = DEFF$. The appropriate *DEFF* value for the F1-F2 subsamples is probably somewhere between 1 (no design effect) and 3 (the median *DEFF* for all variables in the F1-F2 panel sample). I generally take a conservative approach to standard error estimation, using $DEFF = 3$ for all tests of statistical significance unless otherwise specified. In all tests of statistical significance, a critical value of .05 was applied.

References

Becker, B. J. (1990). Coaching for the Scholastic Aptitude Test: further synthesis and appraisal. Review of Educational Research 60(3): 373-417.

Breen, R. (1996). Regression Models: Censored, Sample Selected or Truncated Data. Thousand Oaks, SAGE Publications.

Briggs, Derek C. (in press) Evaluating SAT coaching: gains, effects and self-selection. In: Rethinking the SAT: Perspectives Based on the November 2001 Conference at the University of California, Santa Barbara, R. Zwick, ed., RoutledgeFalmer.

Briggs, D. C. (2001). The Effect of Admissions Test Preparation: Evidence from NELS:88. Chance 14(1): 10-18.

Dorans, N. J. (2002). Recentering and realigning the SAT score distributions: how and why. Journal of Educational Measurement 39(1): 59-84.

Dyer, H. S. (1953). Does Coaching Help? The College Board Review 19: 331-335.

Freedman, D. (1987). As others see us: a case study in path analysis (with discussion). Journal of Educational Statistics 12(101-223).

Freedman, D. (1995). Some issues in the foundations of statistics (with discussion). Foundations of Science 1, 19-83.

Freedman, D. (2002). On specifying graphical models for causation, and the identification problem (Technical Report 601). Berkeley: University of California, Berkeley, Department of Statistics.

Goldberger, A. (1983). Abnormal selection bias. In S. Karlin, T. Amemiya & L. Goodman (Eds.), Studies in econometrics, time series and multivariate statistics. New York: Academic Press.

Greene, W. (1981). Sample selection bias as a specification error: comment. Econometrica 49, 795-798.

Greene, W. H. (1993). Econometric Analysis. New York, Macmillan Publishing Company.

Heckman, J. (1979). Sample selection bias as a specification error. Econometrica 47: 153-161.

Heckman, J. (1978). Dummy endogenous variables in a simultaneous equations system. Econometrica 46, 931-961.

Heckman, J. and Robb, R. (1986). Alternative methods for solving the problem of selection bias in evaluating the impact of treatments on outcomes. In H. Wainer (Ed.), Drawing Inferences from Self-Selected Samples (pp. 63-107). Mahwah, NJ: Lawrence Erlbaum Associates.

Heckman, J. J. and Hotz, J. (1989). Choosing among alternative nonexperimental methods for estimating the impact of social programs: the case of manpower training (with comments). Journal of the American Statistical Association, 84, 862-880.

Holland, P. W. (1986). Statistics and Causal Inference. Journal of the American Statistical Association 81(396), 945-960.

Holland, P.W. (1988). Causal inference, path analysis and recursive structural equations models. In C. Clogg (Ed.), Sociological Methodology (pp. 449-484).

Holland, P. W. (2001). The causal interpretation of regression coefficients. In M. C. Galavotti, P. Suppes, & D. Costantini (Eds.), Stochastic Causality (pp. 173-187): CSLI Publications.

Johnson, N., and S. Kotz. (1971). Distributions in Statistics—Continuous Univariate Distributions, Vol. 2. New York: Wiley.

Jöreskog, K. and D. Sörbom (1996). LISREL 8: User's Reference Guide. Chicago, Scientific Software International.

Lalonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. The American Economic Review 76(4), 604-620.

Little, R. (1985). A note about models for selectivity bias. Econometrica 53(6), 1469-1474.

Little, R. J. and Rubin, D. B. (1987). Statistical analysis with missing data. New York: Wiley.

Lawrence, I., Rigol, G., Van Essen, T. and Jackson, C. (2001). A historical perspective on the SAT. Rethinking the use of the SAT in University Admissions Conference, Santa Barbara, CA. November 17-18, 2001.

Lemann, N. (1999). The big test : the secret history of the American meritocracy. New York, Farrar Straus and Giroux.

Messick, S. (1980). The effectiveness of coaching for the SAT: review and reanalysis of research from the fifties to the FTC. Princeton, Educational Testing Service: 135.

Messick, S. and A. Jungeblut (1981). Time and method in coaching for the SAT. Psychological Bulletin 89: 191-216.

Powers, D. E. (1993). Coaching for the SAT: A summary of the summaries and an update. Educational Measurement: Issues and Practice(Summer): 24-39.

Powers, D. E. and Rock, D. A. (1999). Effects of Coaching on SAT I: Reasoning Test Scores. Journal of Educational Measurement 36(2): 93-118.

Rock, D. (2002). Personal communication. July 24, 2002.

Rosenbaum, P. R. (1995). Observational Studies. New York, Springer-Verlag.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika 70(1): 41-55.

Rosenbaum, P. R. and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. Journal of the American Statistical Association 79, 516-524.

Sesnowitz, M., Bernhardt, K. and Knain, D. M. (1982). An analysis of the impact of commercial test preparation on SAT scores. American Educational Research Journal 19(3): 429-441.

Schwartz, T. (1999). The test under stress. The New York Times. New York: Section 6, Page 30, Column 1.

United States Department of Education (1995). NELS Second Follow-up Student Component Data File User's Manual. Washington, D.C., National Center for Educational Statistics. Available on the world wide web at http://www.nces.ed.gov/surveys/nels88/.

Statacorp. (2001). Stata reference manual set, vol. 4. Stata Press. (www.stata.com).

Vella, F. (1998). Estimating models with sample selection bias: a survey. The Journal of Human Resources 33(1), 127-169.

Wainer, H. ed. (1986). Drawing Inferences from Self-Selected Samples. Mahwah, NJ: Lawrence Erlbaum Associates.

Zuman, J. P. (1988). The effectiveness of special preparation for the SAT: An evaluation of a commercial coaching school. Doctoral dissertation, Harvard University.

Zwick, R. (2002). Fair game? the use of standardized admissions tests in higher education. New York: RoutledgeFalmer.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

TM035049

I. DOCUMENT IDENTIFICATION:

Title: CAUSAL INFERENCE AND THE HECKMAN MODEL	
Author(s): DEREK C BRIGGS	
Corporate Source: UNIVERSITY OF CALIFORNIA, BERKELEY	Publication Date: APRIL 2003

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

[X]

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

[ ]

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

[ ]

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, please

Signature: [signature]	Printed Name/Position/Title: DEREK BRIGGS / Assistant Professor	
Organization/Address: School of Education, University of Colorado Boulder, CO 80309-0249	Telephone: 303-492-6939	FAX: 303-492-7090
	E-Mail Address: derek.briggs@colorado.edu	Date: 6/13/03

(Over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
UNIVERSITY OF MARYLAND
1129 SHRIVER LAB
COLLEGE PARK, MD 20742-5701
ATTN: ACQUISITIONS

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfacility.org